# What is machine learning?

**Machine learning** is the subfield of computer science that gives "**computers the ability to learn without being explicitly programmed.**"

**Arthur Samuel**
American pioneer in the field of computer gaming and artificial intelligence, coined the term "machine learning" in 1959 while at IBM.

# Major machine learning techniques

- Anomaly detection
- Sequence mining
- Dimension Reduction
- Recommendation systems

# Major machine learning techniques

- Regression/Estimation
  - Predicting continuous values
- Classification
  - Predicting the item class/category of a case
- Clustering
  - Finding the structure of data; summarization
- Associations
  - Associating frequent co-occurring items/events

# Major machine learning techniques

- Anomaly detection
  - Discovering abnormal and unusual cases
- Sequence mining
  - Predicting next events; click-stream (Markov Model, HMM)
- Dimension Reduction
  - Reducing the size of data (PCA)
- Recommendation systems
  - Recommending items

# Difference between artificial intelligence, machine learning, and deep learning

- AI components:
  - Computer Vision
  - Language Processing
  - Creativity
  - Etc.
- Machine learning:
  - Classification
  - Clustering
  - Neural Network
  - Etc.
- Revolution in ML:
  - Deep learning

Deep learning involves a deeper level of automation in comparison with most machine learning algorithms

# More about scikit-learn

- Free software machine learning library
- Classification, Regression and Clustering algorithms
- Works with NumPy and SciPy
- Great documentation
- Easy to implement

# scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)
```

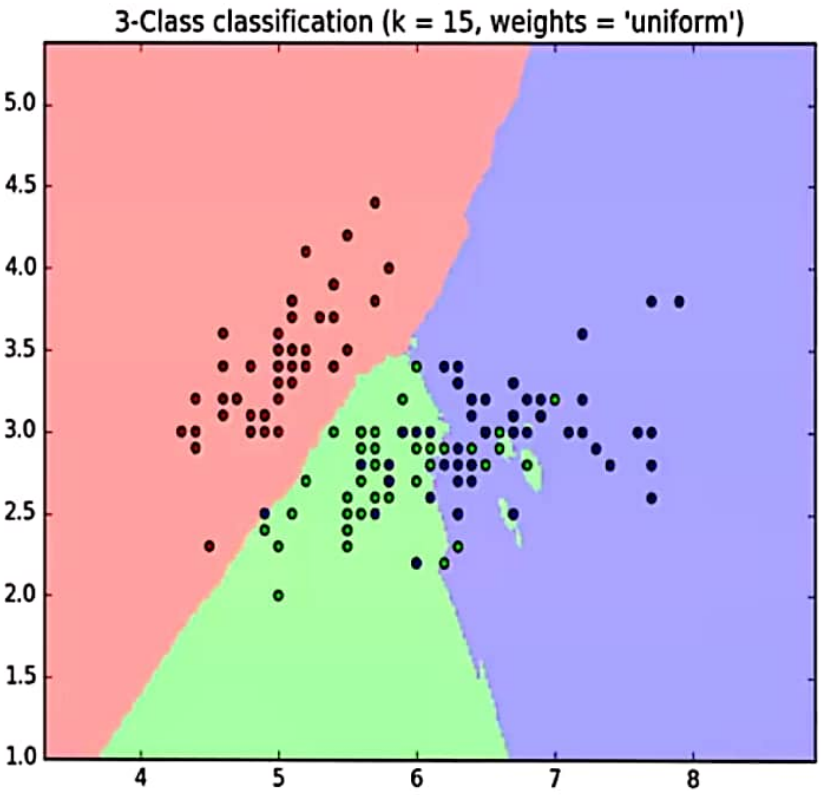
```
clf.fit(X_train, y_train)
```

```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

```
import pickle
s = pickle.dumps(clf)
```

# What is supervised learning?

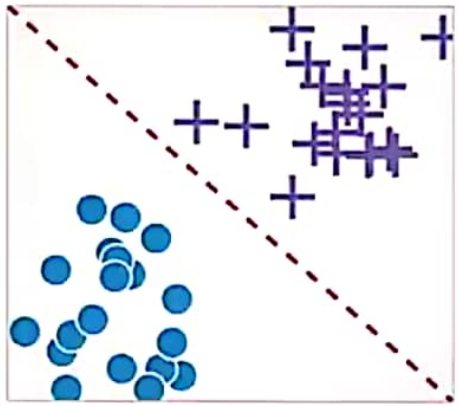



We “teach the model,” then with that knowledge, it can predict unknown or future instances.

# What is classification?

**Classification** is the process of predicting discrete class labels or categories.

| ID | Clump | UnifSize | UnifShape | MargAdh | SingEpiSize | BareNuc | BlandChrom | NormNucl | Mit | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | malignant |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1017122 | 8 | 10 | 10 | 8 | 7 | 10 |  | 7 | 1 | malignant |
| 1018099 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | benign |
| 1018561 | 2 | 1 | 2 | H | 2 | 1 | 3 | 1 | 1 | benign |
| 1033078 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | benign |
| 1033078 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | benign |

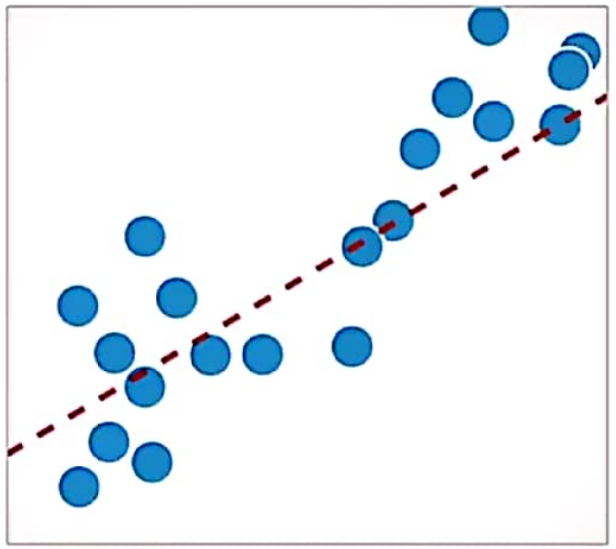


# What is regression?

**Regression** is the process of predicting continuous values.

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |

Continuous Values



# What is unsupervised learning?

| Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Address | DebtIncomeRatio |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | NBA001 | 6.3 |
| 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | NBA021 | 12.8 |
| 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | NBA013 | 20.9 |
| 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | NBA009 | 6.3 |
| 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | NBA008 | 7.2 |
| 6 | 40 | 1 | 23 | 81 | 0.998 | 7.831 | NBA016 | 10.9 |
| 7 | 38 | 2 | 4 | 56 | 0.442 | 0.454 | NBA013 | 1.6 |
| 8 | 42 | 3 | 0 | 64 | 0.279 | 3.945 | NBA009 | 6.6 |
| 9 | 26 | 1 | 5 | 18 | 0.575 | 2.215 | NBA006 | 15.5 |
| 10 | 47 | 3 | 23 | 115 | 0.653 | 3.947 | NBA011 | 4 |
| 11 | 44 | 3 | 8 | 88 | 0.285 | 5.083 | NBA010 | 6.1 |
| 12 | 34 | 2 | 9 | 40 | 0.374 | 0.266 | NBA003 | 1.6 |

The model works on its own to discover information.

# What is unsupervised learning?

| Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Address | DebtIncomeRatio |
|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | NBA001 | 6.3 |
| 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | NBA021 | 12.8 |
| 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | NBA013 | 20.9 |
| 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | NBA009 | 6.3 |
| 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | NBA008 | 7.2 |
| 6 | 40 | 1 | 23 | 81 | 0.998 | 7.831 | NBA016 | 10.9 |
| 7 | 38 | 2 | 4 | 56 | 0.442 | 0.454 | NBA013 | 1.6 |
| 8 | 42 | 3 | 0 | 64 | 0.279 | 3.945 | NBA009 | 6.6 |
| 9 | 26 | 1 | 5 | 18 | 0.575 | 2.215 | NBA006 | 15.5 |
| 10 | 47 | 3 | 23 | 115 | 0.653 | 3.947 | NBA011 | 4 |
| 11 | 44 | 3 | 8 | 88 | 0.285 | 5.083 | NBA010 | 6.1 |
| 12 | 34 | 2 | 9 | 40 | 0.374 | 0.266 | NBA003 | 1.6 |

Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering

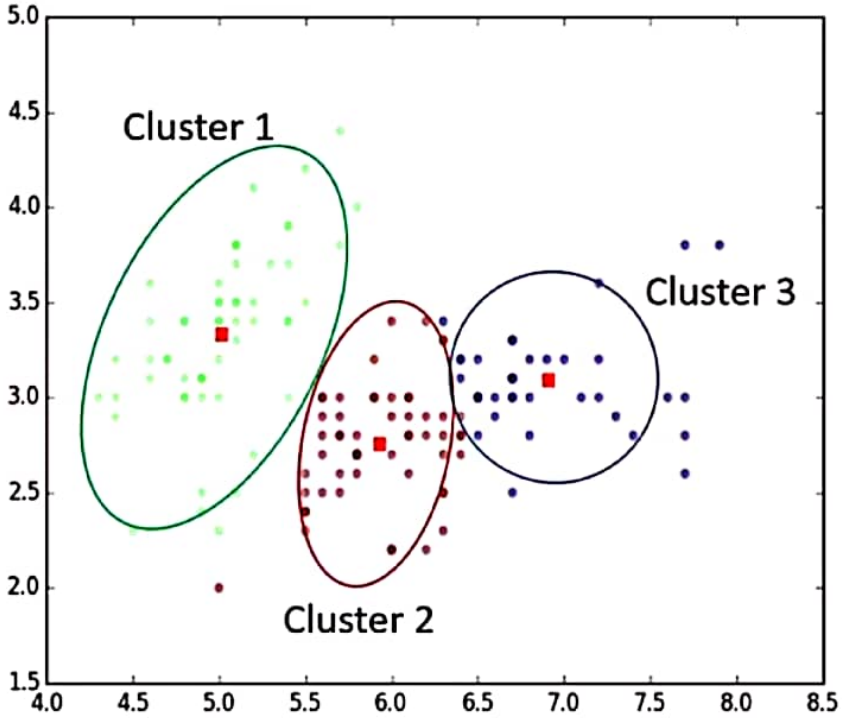
ALL OF THIS DATA IS UNLABELED

The model works on its own to discover information.

# What is clustering?

**Clustering** is grouping of data points or objects that are somehow similar by:

- Discovering structure
- Summarization
- Anomaly detection



# Supervised vs unsupervised learning

## Supervised Learning

- **Classification:**
  Classifies labeled data
- **Regression:**
  Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

## Unsupervised Learning

- **Clustering:**
  Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment

# Regression algorithms

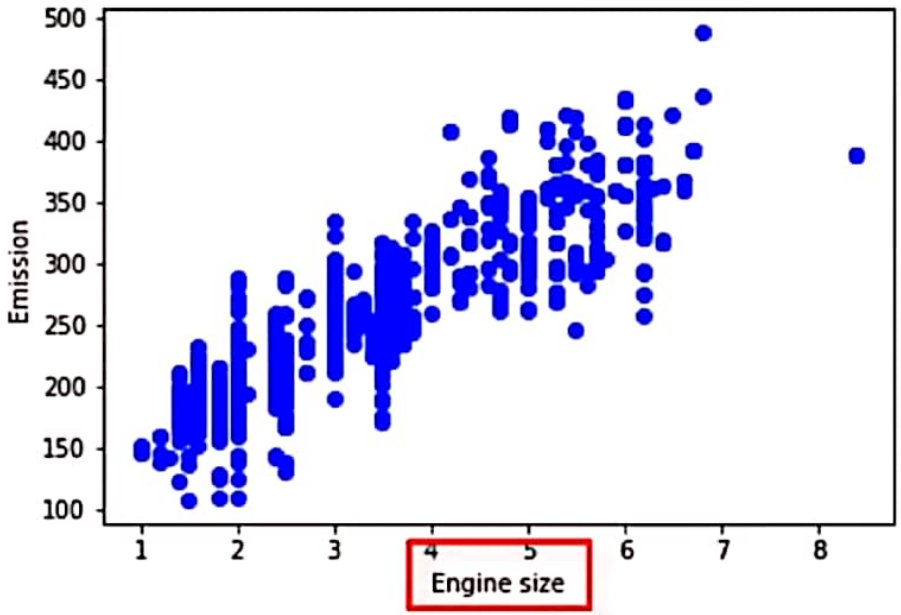
- Ordinal regression
- Poisson regression
- Fast forest quantile regression
- Linear, Polynomial, Lasso, Stepwise, Ridge regression
- Bayesian linear regression
- Neural network regression
- Decision forest regression
- Boosted decision tree regression
- KNN (K-nearest neighbors)

# Using linear regression to predict continuous values

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |

# How does linear regression work?

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |



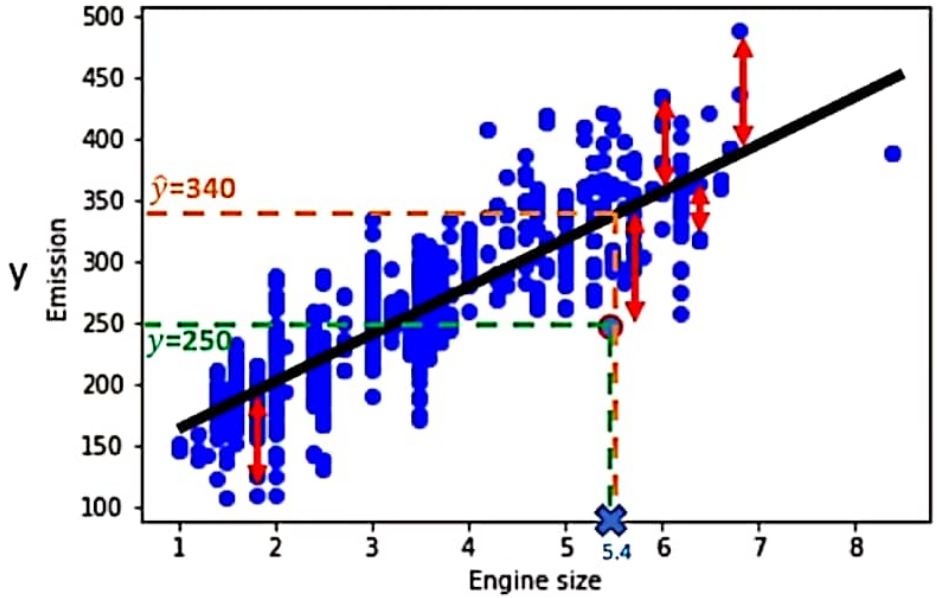
# How to find the best fit?

$x_1$ = 5.4 independent variable
y= 250 actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y}$ = 340 the predicted emission of x1

Error = y- $\hat{y}$
= 250 – 340
= -90

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



# Estimating the parameters

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |

$X_1$ = ENGINESIZE column; $y$ = CO2EMISSIONS column

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ... )/9 = 3.0β

$\bar{y}$ = (196 + 221 + 136 + ... )/9 = 226.22

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

# Estimating the parameters

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |

$X_1$ = ENGINESIZE column; $y$ = CO2EMISSIONS column

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = (2.0 + 2.4 + 1.5 + ... )/9 = 3.03

$\bar{y}$ = (196 + 221 + 136 + ... )/9 = 226.22

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \ldots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \ldots}$$

$$\theta_1 = 39$$

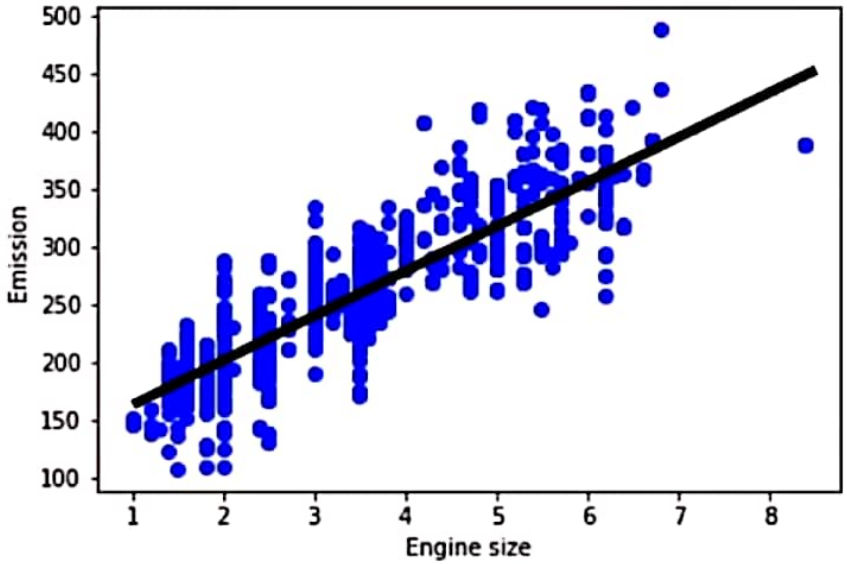
$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 226.22 - 39 * 3.03$$

$$\theta_0 = 125.74$$

# Pros of linear regression

- Very fast
- No parameter tuning
- Easy to understand, and highly interpretable



# Model evaluation approaches

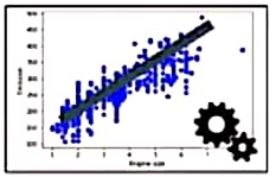
- Train and Test on the Same Dataset
- Train/Test Split

# Best approach for most accurate results?

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | 212 |

Train

Test



| | Prediction |
|---|---|
| 6 | 234 |
| 7 | 256 |
| 8 | 267 |
| 9 | 210 |

Actual values

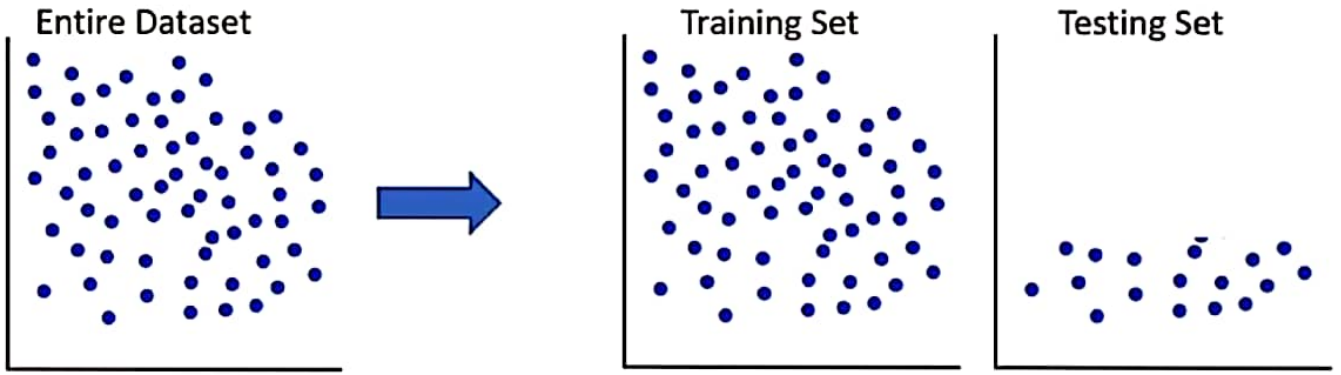
# What is training & out-of-sample accuracy?

- **Training Accuracy**
  - High training accuracy isn't necessarily a good thing
  - Result of over-fitting
    - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model

- **Out-of-Sample Accuracy**
  - It's important that our models have a high, out-of-sample accuracy
  - How can we improve out-of-sample accuracy?

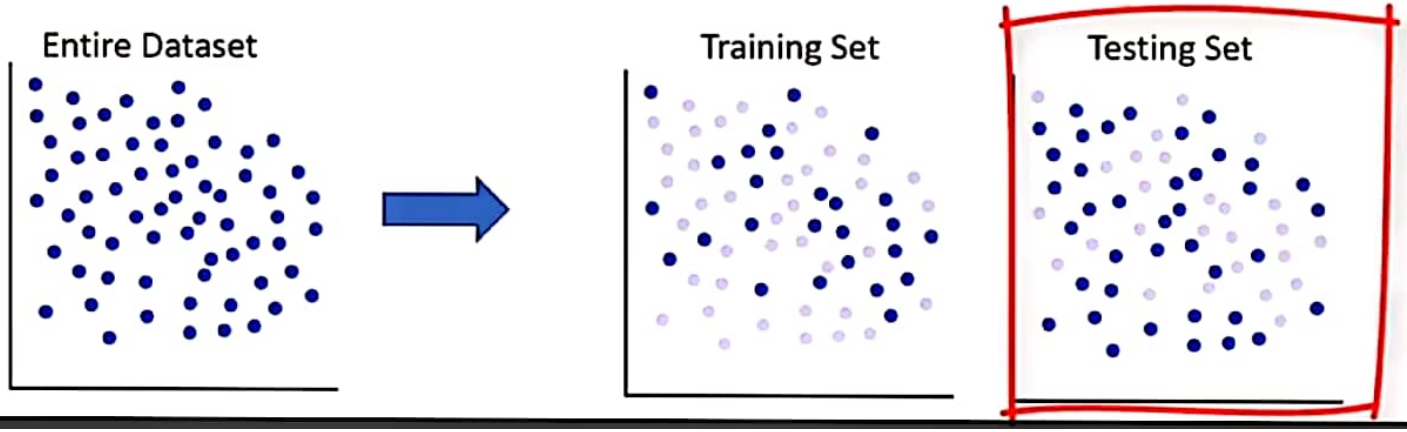
# Train/Test split evaluation approach

**Test on a portion of train set**



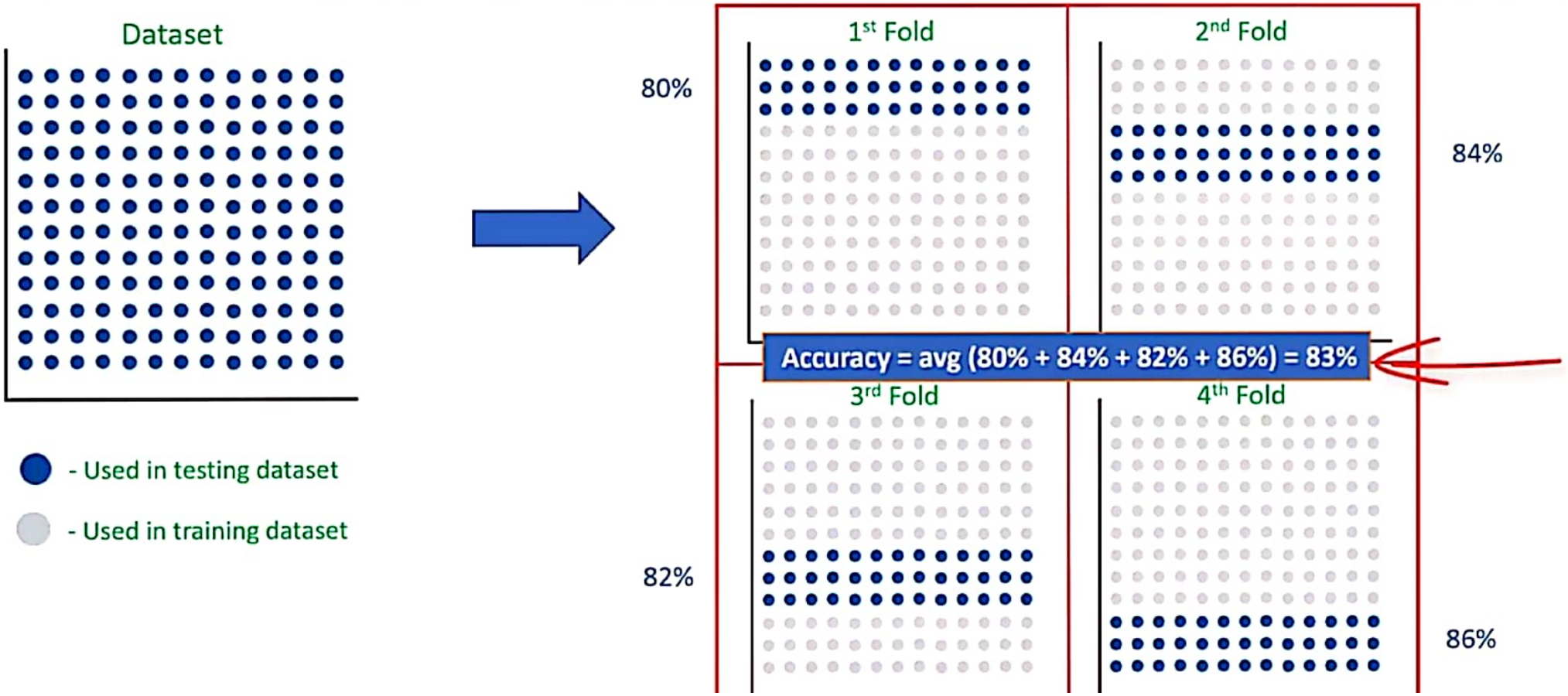

- Test-set is a portion of the train-set
- High "training accuracy"
- Low "out-of-sample accuracy"

**Train/Test Split**




- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy

# How to use K-fold cross-validation?



# Regression accuracy

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | 212 |

Test

$y$

Actual values

$$Error = \frac{(232 - 234) + (255 - 256) + \ ...}{4}$$

$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

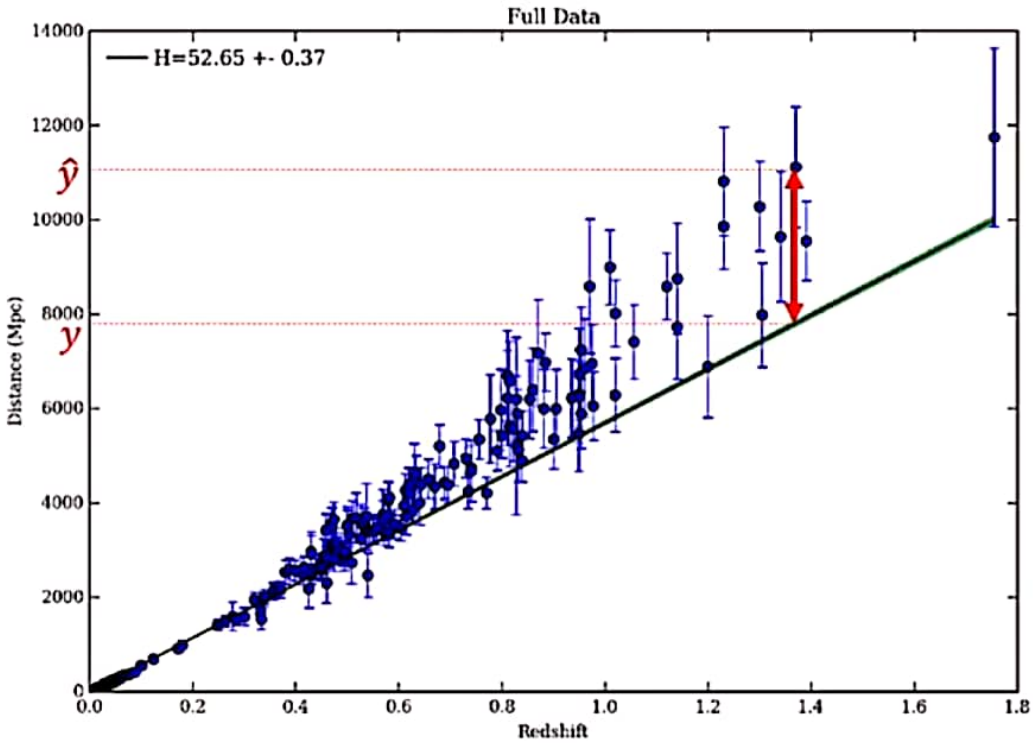
- MAE
- MSE
- RMSE
- ...

$\hat{y}$

| | Prediction |
|---|---|
| 6 | 234 |
| 7 | 256 |
| 8 | 267 |
| 9 | 210 |

Predicted values

# What is an error of the model?




$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

# Examples of multiple linear regression

- Independent variables effectiveness on prediction
  - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?

- Predicting impacts of changes
  - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?

# Predicting continuous values with multiple linear regression

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ... ]$$

X: Independent variable

Y: Dependent variable

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |

# Using MSE to expose the errors in the model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of $x_i$

$y_i = 196$ actual value of $x_i$

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSIONS |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |

# Estimating multiple linear regression parameters

- How to estimate $\theta$?
  - Ordinary Least Squares
    - Linear algebra operations
    - Takes a long time for large datasets (10K+ rows)
  - An optimization algorithm
    - Gradient Descent
    - Proper approach if you have a very large dataset

# Q&A – on multiple linear regression

- How to determine whether to use simple or multiple linear regression?
- How many independent variables should you use?
- Should the independent variable be continuous?
- What are the linear relationships between the dependent variable and the independent variables?

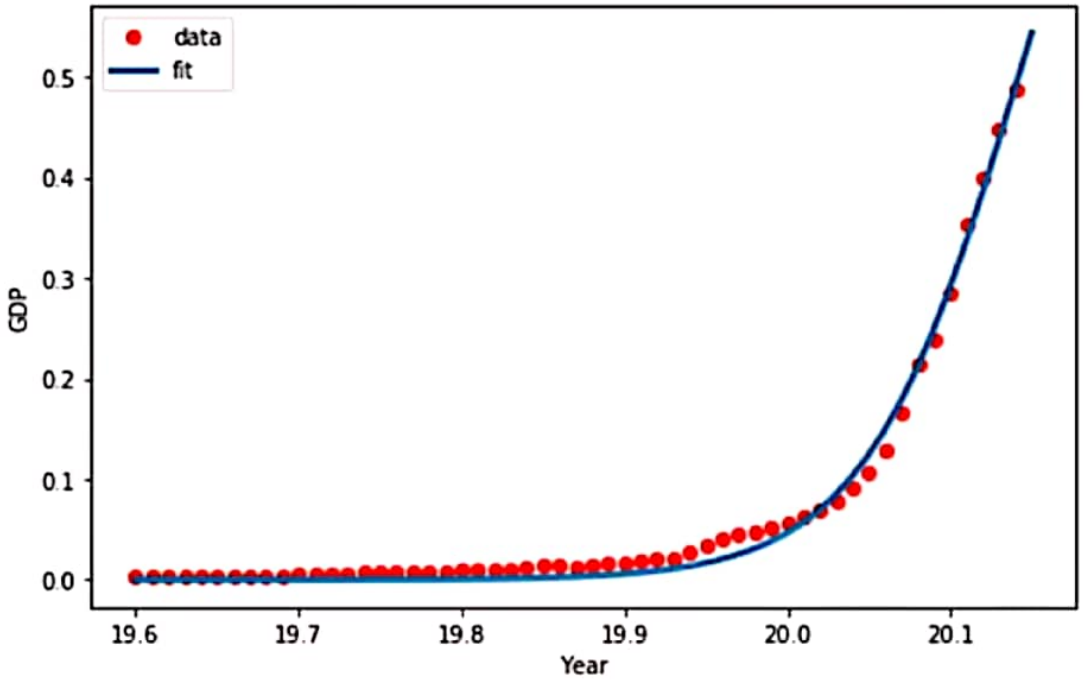
# 11 Characteristics of Truly Inspiring Leaders

1. They have great positive energy.
2. They truly value everyone on their team.
3. They have a clear vision for the future.
4. They listen.
5. They communicate effectively.
6. They are very trustworthy.
7. They are passionate.
8. They practice humility and empathy
9. They are authentic
10. They have a sense of purpose
11. They have a great sense of humor

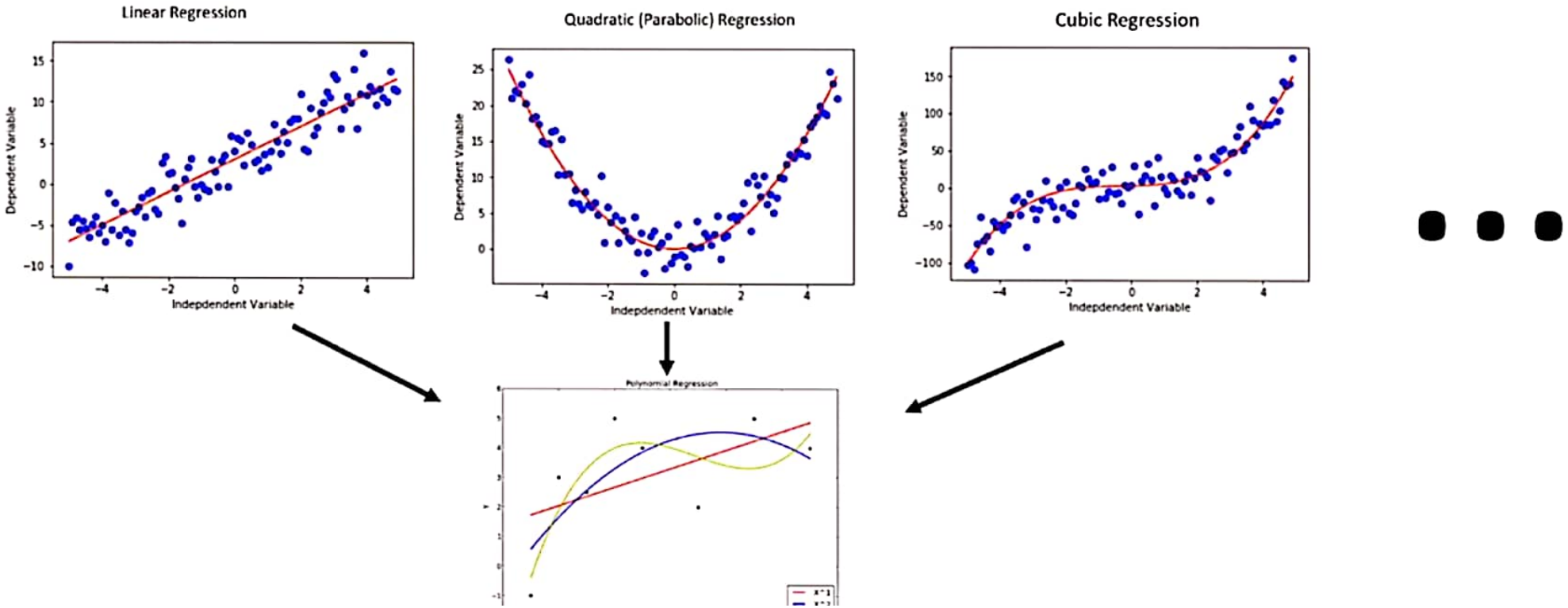
Leadership First
Believe In The Impossible

# Should we use linear regression?

| | Year | Value |
|---|---|---|
| 0 | 1960 | 5.918412e+10 |
| 1 | 1961 | 4.955705e+10 |
| 2 | 1962 | 4.668518e+10 |
| 3 | 1963 | 5.009730e+10 |
| 4 | 1964 | 5.906225e+10 |
| 5 | 1965 | 6.970915e+10 |
| 6 | 1966 | 7.587943e+10 |
| 7 | 1967 | 7.205703e+10 |
| 8 | 1968 | 6.999350e+10 |
| 9 | 1969 | 7.871882e+10 |
| ... | ..... | ..... |



# Different types of regression



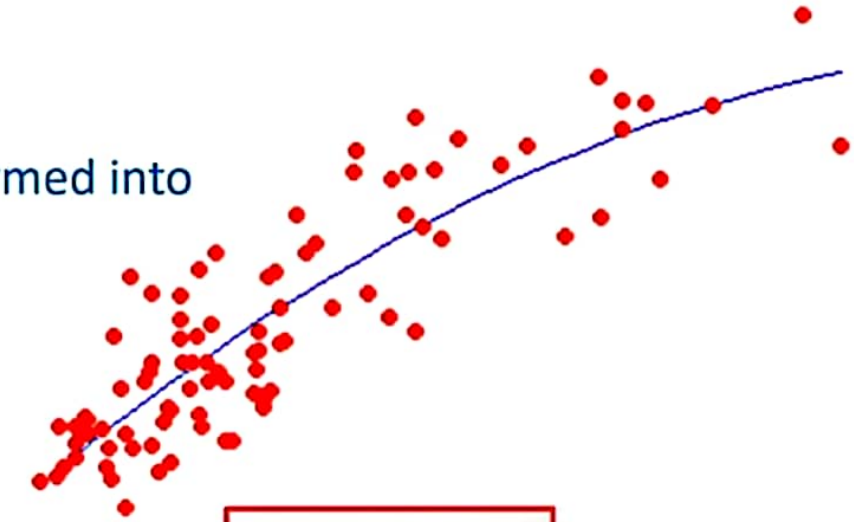
# What is polynomial regression?

- Some curvy data can be modeled by a polynomial regression
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- A polynomial regression model can be transformed into linear regression model.

$$x_1 = x$$
$$x_2 = x^2$$
$$x_3 = x^3$$



$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$ → Multiple linear regression → Least Squares

Minimizing the sum of the squares of the differences between $y$ and $\hat{y}$

# What is non-linear regression?

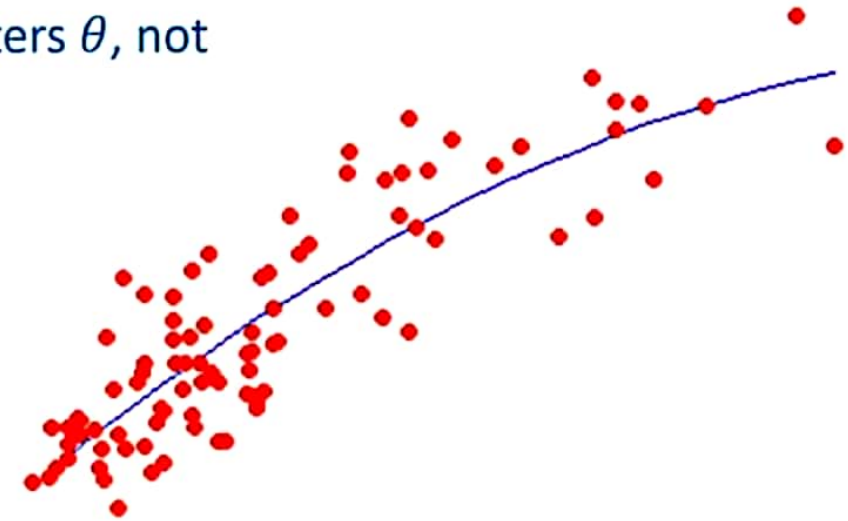
- To model non-linear relationship between the dependent variable and a set of independent variables
- $\hat{y}$ must be a non-linear function of the parameters $\theta$, not necessarily the features x

$$\hat{y} = \theta_0 + \theta_2{}^2 x$$

$$\hat{y} = \theta_0 + \theta_1 \theta_2{}^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_1{}^{(x-\theta_2)}}$$



# Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
  - Inspect visually
  - Based on accuracy

- How should I model my data, if it displays non-linear on a scatter plot?
  - Polynomial regression
  - Non-linear regression model
  - Transform your data